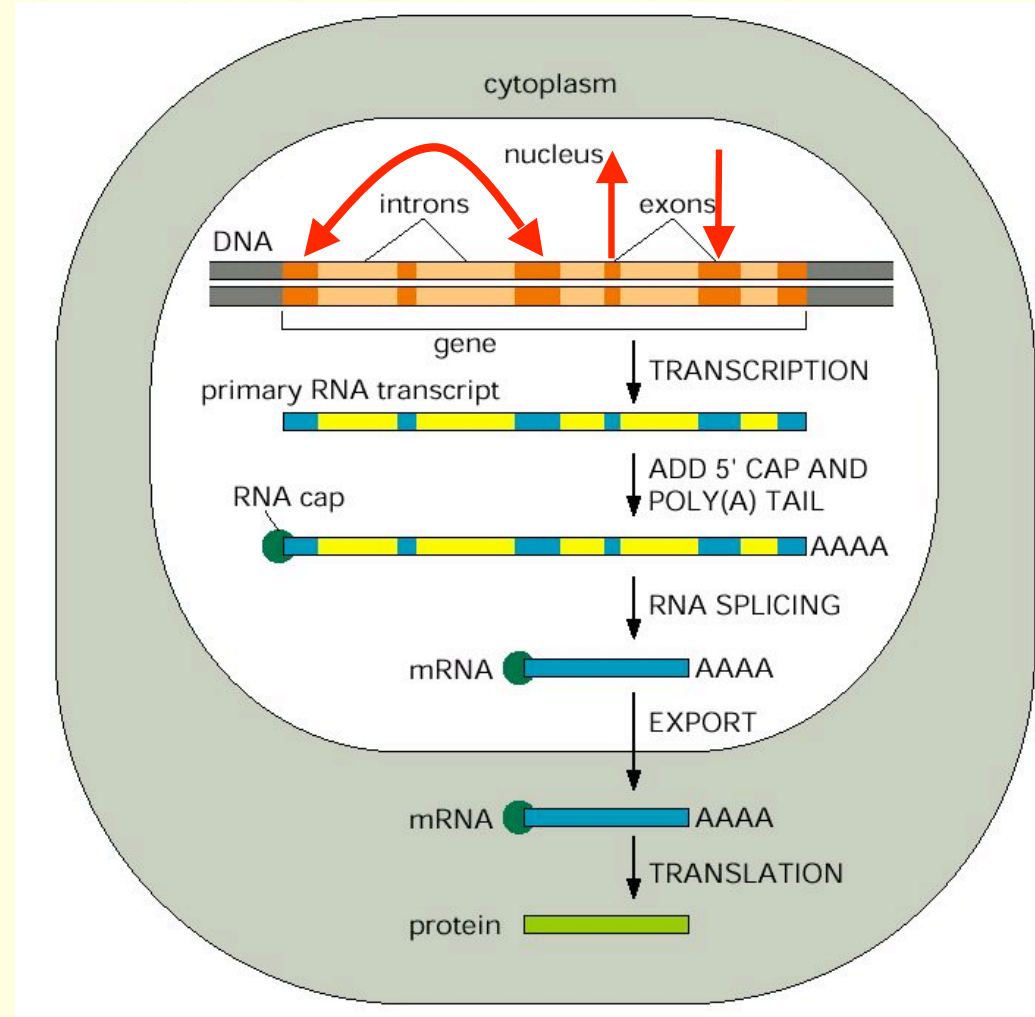
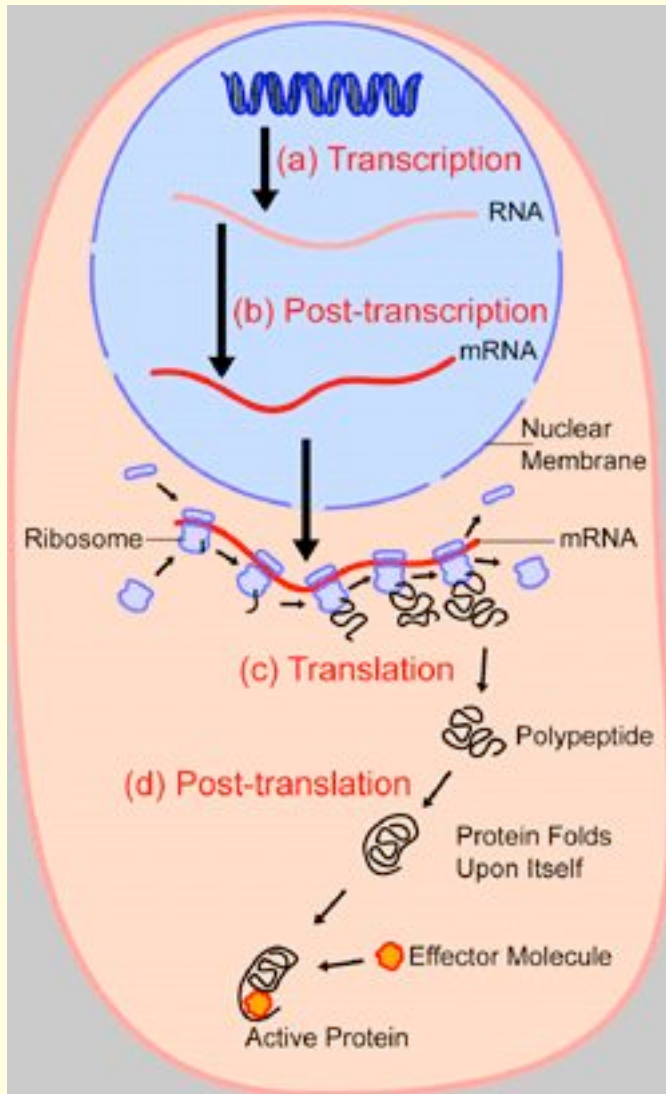


ECE697S: Topics in Computational Biology

Lecture 3: Sequence Alignment Methods



Organisms evolve through mutation events that modify sequence.

Types of Mutations

- DNA can be modified by:
 - insertions, deletions, substitutions
 - repeats, rearrangements
- 1.5% of mammalian DNA codes for proteins, 5-7% is functional.
- We are interested in how functional DNA is modified, and at what rate.

Sequence Alignment

- We'd like to measure the “similarity” between a given pair, or set, of sequences.
- We can model a sequence by a string from a 4- or 20-character alphabet.

Dynamic Programming

- Problems that can be solved by dynamic programming have a *local optimality* property.
- Dynamic Programming usually involves generating a table of costs of various subproblems, and finding the optimal combination of costs.

Edit Distance

- A **string** is a sequence of characters drawn from some alphabet.
- $D(s_1, s_2)$ = the minimum number of **edits** needed to convert **string** s_1 into **string** s_2 .
- An **edit** is considered to be an insertion, deletion, or substitution.

Edit Distance

- Where else do we see insertions, deletions and substitutions?
- The edit distance between two gene sequences can be viewed as a type of *global* alignment.
- We can find the actual alignment by backtracking.

Dynamic programming matrix:

		j → (sequence y)								
		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
i ↓ (sequence x)	0	0	-6	-12	-18	-24	-30	-36	-42	-48
	1 T	-6	5	-1	-7	-13	-19	-25	-31	-37
	2 T	-12	-1	3	-3	-2	-8	-14	-20	-26
	3 C	-18	-7	-3	8	2	3	-3	-9	-15
	4 A	-24	-13	-9	2	6	0	1	-5	-4
	5 T	-30	-19	-15	-4	7	4	-2	6	0
	M = 6 A	-36	-25	-21	-10	1	5	2	0	11

Optimum alignment scores 11:

T	-	-	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	-6	-6	+5	+5	-2	+5	+5

Levenshtein defined edit distance, but did not give an algorithm (1964). The dynamic programming algorithm is “folklore”, the application to global sequence alignment is attributed to Needleman-Wunsch (1970).

Problems on Sequences

- Longest Common Subsequence is NP-complete for k sequences, but for 2 sequences, we can solve the problem using Edit Distance.
- We can modify the penalties for the edits to be “realistic” for DNA or protein sequences.

Global Sequence Alignments

- We can generalize our approach, by having a **scoring matrix** as input.
- DNA is simpler to model than amino acids: penalties are based on empirical observations (PAM, BLOSUM).

Local Alignment

- What if we are interested in conserved regions (i.e. translocations/repeats) instead?
- Smith-Waterman (1981) observed that a simple modification to the global alignment algorithm works.
- Both local and global alignment algorithms are not good for multiple sequences.

Multiple Sequence Alignment

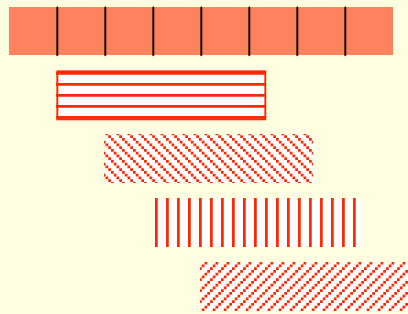
- Applying the DP algorithm to a set of k sequences yields a running time of roughly $O((2n)^k)$.
- We must resort to heuristics...

FASTA

- A heuristic approximation to Smith-Waterman, for multiple sequences.
- First, identify similar sequences from a database, i.e., rank by word agreement.
- Then, align substrings corresponding to maximally agreeing regions using Smith-Waterman.

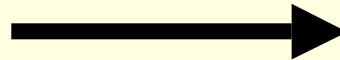
BLAST

Query Sequence

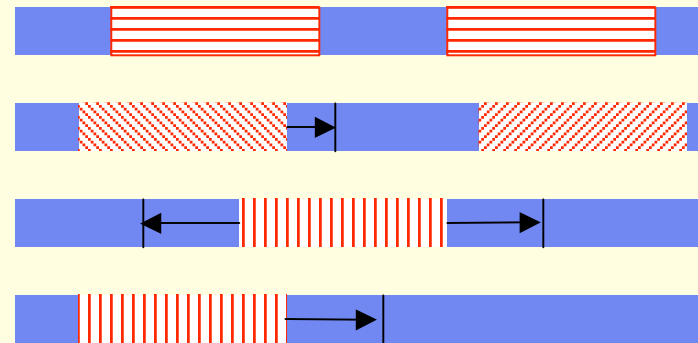


words of length w

Identify hits,
then MSPs



Sequence Library



- Maximum Segment Pair: 2 substrings of equal length with best alignment score.
- Can be extended to handle gaps and match position (Gapped-BLAST, PSI-BLAST).

Genomic Rearrangement

- Genes are organized into **synteny blocks** on a chromosome, which can be rearranged between organisms.
- In unichromosomal genomes, the most common event is a **reversal**.
- The **Reversal Distance Problem**: Find shortest sequence of reversals that sorts a permutation.

Sorting Permutations

- Two sequences of synteny block represents two permutations.
- How many reversals occurred?
- The reversal distance problem is equivalent to “sorting” a permutation.

A Simple Algorithm

- Greedily choose reversal that correctly places permutation indices.
- Works, but approximation ratio is poor: $(n-1)/2$.
- Reversal Distance is NP-complete!

Approximation Algorithms

- If finding the minimum reversal distance is NP-complete, can we approximate the reversal distance?
- For example, can we get within 10% of the reversal distance, in time polynomial in the size of the permutation?

Breakpoints

- Let a **breakpoint** be pair of adjacent elements that are out of order. The number of breakpoints is at most twice the number of reversals.
- Eliminating all breakpoints is equivalent to sorting our permutation.
- Greedily eliminate breakpoints.

A Better Algorithm

- Define a strip as an increasing or decreasing portion of a permutation.
- **Theorem:** If a permutation has an increasing strip, there is a reversal that decreases the number of breakpoints.
- **Theorem** (Kececioglu/Sankoff 1995): The greedy approach is a 4-approximation to reversal distance.

Algorithms for Reversal Distance

Distance

- The best algorithm for reversal distance has an approximation ratio of 1.35 (Berman, Hannenhalli, Karpinski 2002).
- A signed permutation has elements between breakpoints alternate in sign.
- If we have signed permutations, then we can solve the reversal distance problem in $O(n^2)$ time (Kaplan, Shamir, Tarjan 1997).